FINNEGAN, HENDERSON, FARABOW, GARRETT & DUNNER, L.L.P.

1300 I STREET, N. W.
WASHINGTON, DC 20005-3315

202 • 408 • 4000
FACSIMILE 202 • 408 • 4400

ATLANTA
404•653•6400

PALO ALTO
650•849•6600

WRITER'S DIRECT DIAL NUMBER:

(202) 408-6070

TOKYO
011•813•3431•6943

BRUSSELS
011•322•646•0353

September 30, 1999

ATTORNEY DOCKET NO. 01413.0010

Box PATENT APPLICATION
Assistant Commissioner for Patents
Washington, D.C. 20231

Re: New U.S. Patent Application
       Title: **METHOD AND APPARATUS FOR EXTRACTING**
       ATTRIBUTES FROM SEQUENCE STRINGS AND
       BIOPOLYMER MATERIAL
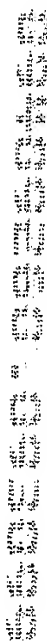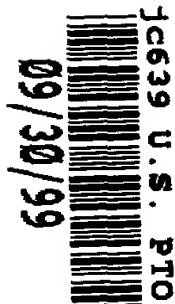   <u>Inventors: Jeffrey D. Saffer et al.</u>

Sir:

        We enclose the following papers for filing in the United States Patent and Trademark Office in connection with the above patent application.

1.   Application - 30 pages, including 8 independent claims and 21 claims total.

2.   Drawings - 7 sheets of drawings (Figures 1-7).

3.   Please charge $1172 to our deposit account number 06-0196 representing a $760 filing fee and $412 for additional claims.

        This application is being filed under the provisions of 37 C.F.R. § 1.53(f). Applicants await notification from the Patent and Trademark Office of the time set for filing the Declaration.

        Please accord this application a serial number and filing date.

The Commissioner is hereby authorized to charge any additional filing fees due and any other fees due under 37 C.F.R. § 1.16 or § 1.17 during the pendency of this application to our Deposit Account No. 06-0916.

Respectfully submitted,

FINNEGAN, HENDERSON, FARABOW,
GARRETT & DUNNER, L.L.P.

By:_____
James T. Wilson
Reg. No. 41,439

JTW/jk
Enclosures

UNITED STATES PATENT APPLICATION

OF

Jeffrey D. SAFFER

Augustin J. CALAPRISTI

Nancy E. MILLER

Randall E. SCARBERRY

Heidi J. SOFIA

Lisa C. STILLWELL

Guang CHEN

and

Phillip J. MONROE

FOR

METHOD AND APPARATUS FOR EXTRACTING ATTRIBUTES FROM SEQUENCE STRINGS AND BIOPOLYMER MATERIALS

LAW OFFICES
FINNEGAN, HENDERSON,
FARABOW, GARRETT,
& DUNNER, L.L.P.
1300 I STREET, N. W.
WASHINGTON, D.C.20005
202-408-4000

RELATED APPLICATIONS

The following identified U.S. patent applications are relied upon and are incorporated by reference in their entirety in this application:

U.S. Patent Application Ser. No. 08/695,455, entitled "THREE-DIMENSIONAL DISPLAY OF DOCUMENT SET," filed on August 12, 1996;

U.S. Patent Application Ser. No. 08/713,313, entitled "SYSTEM FOR INFORMATION DISCOVERY," filed on September 13, 1996;

U.S. Patent Application Ser. No. _____, entitled "METHODS AND APPARATUS FOR DISPLAYING DISPARATE TYPES OF INFORMATION USING AN INTERACTIVE SURFACE MAP," filed on the same date herewith by Jeffrey Saffer, et al.; and

U.S. Patent Application Ser. No. _____, entitled "DATA PROCESSING, ANALYSIS, AND VISUALIZATION SYSTEM FOR USE WITH DISPARATE DATA TYPES," filed on the same date herewith by Jeffrey Saffer, et al..

BACKGROUND OF THE INVENTION

A. Field of the Invention

The present invention relates to extracting attributes from sequence strings and from information representing biopolymer materials, and more particularly to a method and apparatus which extracts attributes from information representing biopolymer material to create objects useful for analyzing large amounts of data using multivariate analysis.

B. Description of the Related Art

DNA, RNA, and proteins represent key functional units in biological systems. DNA is composed of nucleotide subunits (deoxyadenosine, deoxythymidine, deoxycytidine, and deoxyguanosine) linked together to form an array of biopolymer material. Often, the linked chain is bound to a complementary chain to form a double helix. The code contained within the DNA is of multiple types. Some sequences within the DNA are recognized by regulatory factors and control how the biopolymer information is expressed. Some sequences encode structural attributes that contribute to the overall use of the biopolymer material. And some sequences encode the RNA or proteins that carry out functions within the cell. For simplicity, DNA is usually represented as an ordered string of the deoxynucleotides (e.g., GATTCTAGGA), but that simple string reflects the full function of the molecule. The RNA copy of the DNA is also a chain of nucleotides (adenosine, uridine, cytidine, and guanosine being the major ones) (e.g., AUGGACCAUA). Some RNAs are translated into proteins, which are strings of amino acid building blocks.

There are 20 principal amino acids building blocks, and proteins are often represented simply by an ordered string of sequence letters (e.g., MRKLAGQPS). The function of proteins is not, however, fully contained within this simple string, since the building blocks can be modified in multiple ways within a cell. Nonetheless, the sequence of the amino acids is the primary contributor to the function of the protein.

The realm of bioinformatics is largely focused on trying to predict the function of genomic sequences. This work involves comparing the strings of information (genomic

- 3 -

sequences), functional properties, and behavior of known and unknown entities, thereby providing a basis for predicting the similar function of sequences with similar properties. These methods, however, are not usually geared toward simultaneous analysis of a large number of sequences. Thus, it is difficult to get an overview of how all the unknown and known sequences relate to each other from these methods.

A number of multivariate analysis methods, including those geared toward data visualization and data mining, are available. In each case, a data object is represented as a high-dimensional vector, where the number of dimensions is equal to the number of independent attributes required to describe the data object.

For data strings, such as genome sequences, however, there are relatively few methods that have been applied to represent the information as a high-dimensional vector. One method creates a signature for protein sequences based on the occurrence of all possible amino acid dimers (or pairs of amino acids). See van Heel M., *A New Family of Powerful Multivariate Statistical Sequence Analysis Techniques*, 220 J. Mol Biol 877-887 (1991). Application of this method with 20 amino acids resulted in a 20 x 20 or 400-dimensional representation for each protein for comparison using cluster analysis.

Another method also includes information about individual amino acids (composition) and descriptive information such as length of the sequence and pl (isoelectric point). These composite vectors were then used for searching data sets to identify similar sequences. See Hobohm U. and Sander C., *A Sequence Property Approach to Searching Protein Databases*, 251 J. Mol. Biol. 390-399 (1995).

While the goal in creating vectors in the above methods was to create a surrogate for functional information in the proteins, these methods do not provide sufficient discrimination to represent the subtle differences between most genomic sequences.

A different approach for mathematical representation of sequences for multivariate analysis is to use an ordination method. See Higgins D.G., *Sequence Ordinations: a Multivariate Analysis Approach to Analyzing Large Sequence Data Sets*, 8 Comput. Appl. Biosci. 15-22 (1992). Such a method uses the square root of the percentage difference between two sequences as a Euclidean distance. Then, each protein is represented within a distance matrix derived from all comparisons. The usefulness of percentage differences as a distance measure, however, is limited to closely related sequences. U.S. Patent No. 5,930,784 to Hendrickson, issued July 27, 1999, provides an example of using geometric distances among all items in a data set for data mining.

These methods are quite limited, however, when comparing proteins with limited similarity or when analyzing a large number of proteins simultaneously.

SUMMARY OF THE INVENTION

Systems and methods consistent with the present invention generate a high-dimensional vector representation of a sequence string of units in a data set, including a plurality of sequence strings, by operations for dividing each of respective sequence strings into blocks of three units or more to create a vocabulary of blocks,

defining a respective vector axis to correspond to blocks in the vocabulary, determining for each vector axis whether a string includes a block corresponding to the respective vector axis, and creating a high-dimensional vector based on the determining.

Methods consistent with the present invention generate a high-dimensional vector representation of an item of biopolymer material in a data set including a plurality of items of biopolymer material by selecting predefined domains of the plurality of items of biopolymer materials, defining a respective vector axis to correspond to the selected domains, determining for each vector axis whether an item of biopolymer material includes a domain corresponding to the respective vector axis, and creating a high-dimensional vector based on the determining.

Methods consistent with the present invention generate a high-dimensional vector representation of an item of biopolymer material in a data set including a plurality of items of biopolymer material, by defining each item of biopolymer material in the data set as a surface using descriptors of at least one of structure and function, defining a respective vector axis to correspond to descriptors, determining for each vector axis whether an item of biopolymer material includes a descriptor corresponding to the respective vector axis, and creating a high-dimensional vector based on the determining.

Methods consistent with the present invention generate a high-dimensional vector representation of an item of biopolymer material in a data set including a plurality of items of biopolymer material by comparing information regarding each biopolymer material of the plurality to information regarding each other biopolymer material to

provide a respective result, arranging the results in a square matrix indexed by the plurality of items of biopolymer materials, creating a high-dimensional vector for an item of biopolymer material based on a row or column of the matrix, and creating a distance matrix based on the high-dimensional vector.

An apparatus consistent with the present invention generates a high-dimensional vector representation of an item of biopolymer material in a data set including a plurality of items of biopolymer material and includes at least one memory having program instructions, and at least one processor configured to execute the program instructions to perform the operations to generate the high-dimensional vector.

A computer-readable medium consistent with present invention contains instructions for controlling a computer system to generate a high-dimensional vector representation of an item of biopolymer material in a data set including a plurality of items of biopolymer material.

It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory only and are not restrictive of the invention, as claimed.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate the implementations of the invention and together with the description, serve to explain the principles of the invention.

FIG. 1 is a diagram of an exemplary computing system with which the present invention may be implemented;

FIG. 2 is a flow chart of steps used to visualize a data set of biopolymer material consistent with the present invention;

FIG. 3 is a flow chart of steps of an implementation of a method for creating context vectors for sequence strings;

FIG. 4 is a flow chart of steps of an implementation of a method for creating context vectors for biopolymer material;

FIG. 5 is a flow chart of steps of another implementation of a method for creating context vectors for biopolymer material;

FIG. 6 is a flow chart of steps of another implementation of a method for creating context vectors for biopolymer material; and

Fig. 7 is an illustration of a square matrix created in the method of Fig. 6.

DETAILED DESCRIPTION OF THE INVENTION

Reference will now be made in detail to the construction and operation of an implementation of the present invention which is illustrated in the accompanying

drawings. The present invention is not limited to this implementation but it may be realized by other implementations.

A.   Overview

Methods and systems consistent with the present invention extract attributes from sequence strings and from information representing biopolymer materials. Sequence strings are useful to provide an abstract representation of a complex object. In the biosciences, sequence strings are used to represent biopolymer materials, which are macromolecules found within a living thing, such as proteins, nucleic acids (such as DNA or RNA), polysaccharides, and mulitprotein and other complexes. Using sequence strings, bioscientists are provided with a common language to compare the features of different biopolymer materials.

Nevertheless, sequence strings of biopolymer materials are difficult to analyze, especially in large numbers. For example, most people can only remember seven units of unrelated information at a time (such as a seven-digit phone number) and sequence strings can be much longer than seven digits. With the need for analysis of a large number of biopolymer materials, analysis by a person is impractical. Fortunately, computer processing can make such analysis easier.

Computer processing, however, must be able to understand the sequence string. Particularly, when one seeks to view a relationship between numerous objects having numerous attributes, the computer must be able to define the attributes of the objects.

Accordingly, methods and systems consistent with the present invention provides a computer with understandable attributes of biopolymers and other objects.

Using such methods and systems, relationships of information corresponding to structures such as biopolymer materials can be examined simultaneously. Thus, a user can discover relationships of biopolymer material not based on linear alignments but rather on overall characteristic relationships. Such relationships could be used to create, for example, a cure for a particular disease while eliminating side-effects.

B.   Architecture

Fig. 1 is a diagram of an exemplary computer system 100 that can carry out processes consistent with the present invention. Computer system 100 includes a processor 102 and a memory 104 coupled to processor 102 through a bus 106. Processor 102 fetches computer instructions from memory 104 and executes those instructions. Processor 102 also (1) reads data from and writes data to memory 104, (2) sends data and control signals through bus 106 to one or more computer output devices 120, (3) receives data and control signals through bus 106 from one or more computer input devices 130 in accordance with the computer instructions, and (4) transmits and receives data through bus 106 and router 125 to a network.

Memory 104 can include any type of computer memory including, without limitation, random access memory (RAM), read-only memory (ROM), storage devices that include storage media such as magnetic and/or optical disks, and network-based memory devices. Memory 104 includes a computer process 110, such as a Web

- 10 -

browser or Web server software or a process consistent with the present invention. A computer process includes a collection of computer instructions and data that collectively define a task performed by computer system 100.

Computer output devices 120 can include any type of computer output device, such as a printer 124, and display 122 such as a cathode ray tube (CRT), a light-emitting diode (LED) display, or a liquid crystal display (LCD). Display 122 preferably displays the graphical and textual information received from a computer process. Each of computer output devices 120 receives from processor 102 control signals and data and, in response to such control signals, displays data.

User input devices 130 can include any type of user input device such as a keyboard 132, or keypad, or a pointing device, such as an electronic mouse 134, a trackball, a lightpen, a touch-sensitive pad, a digitizing tablet, thumb wheels, or a joystick. Each of user input devices 130 is used to generate signals in response to physical manipulation by a user and transmits those signals through bus 106.

C.   <u>Architectural Operation</u>

Fig. 2 illustrates steps of a computer process used to visualize a data set representing biopolymer material, e.g. a protein data set. First, the data set is accessed (step 210). A context vector is created for each of the biopolymer materials, e.g. proteins, in the data set (step 220). The context vector will be of a very high-dimension, so as to represent many attributes of the protein. The context vectors are used to visualize the data set to find related or unrelated attributes of the proteins (step 230).

- 11 -

1.   Sequence Data

A first method for creating context vectors for biopolymer material uses sequence data and is shown in Fig. 3.

In this method, each protein in the data set is identified as a respective series of sequence letters corresponding to the amino acid building blocks of the protein. The series of sequence letters lacks ascertainable attributes (step 310). Therefore, is it necessary to use special processing to determine the attributes of the sequence data.

To determine attributes of the sequence data, a vocabulary is created to define attributes of the protein (step 320). Generally, the sequence data is treated as a text document consisting of a number of words, or n-grams. For example, if a series of sequence letters representing a protein is ABCDEF, where each letter is an amino acid, then the protein could be represented as consisting of the n-gram words ABC and DEF, where n equals three. The selection of the length of the n-gram word is important. If a word consists of two or fewer units of the sequence string, such as amino acids, then each word will not convey much information. On the other hand, if words are too long then every word would be unique and, thus, no comparisons of common attributes would be possible. In one implementation, n-grams words of three units (three-mers) and n-gram words of four units (four-mers) provide enough discrimination for proteins, which have a 20-letter alphabet, and create a vocabulary of reasonable size. Nevertheless, the use of one-mers, two-mers, and large n-mers, is within the scope of the invention.

- 12 -

Once a proper word length is determined, the protein sequence is represented by the words. However, protein functionality is often determined by its distinct stretches of its amino acids. Thus, arbitrarily breaking the protein into words might not generate the particular strain of amino acids that best define the protein's function. In order to not disturb the particular strain of amino acids, the protein is preferably divided into a set of overlapping words. For example, the sequence ABCDEF would be represented as ABC, BCD, CDE, and DEF using three-mers instead of simply using ABC and DEF.

Using the process of dividing each series of sequence letters into blocks of overlapping amino acids, a vocabulary of n-gram words is created that indicate the attributes of each of the proteins in the data set.

Referring to Fig. 3, each of the words of the vocabulary is set to a respective axis in a high-dimensional space (step 330). Because of the 20-letter alphabet for proteins, using three-mers will provide a 20^3-dimensional space as a maximum. A large vocabulary, however, can create problems, because, for example, processing the vocabulary will require more computational power. Therefore, the vocabulary, and thus the high-dimensional space, is preferably reduced.

For example, statistical analysis can define the limited number of n-gram words that are most likely to convey information. Such statistical analysis can draw upon existing statistical analysis for natural language documents. For example, in U.S. Patent Application Ser. No. 08/713,313, entitled "System for Information Discovery," words that are too frequent or too infrequent in the data set can be ignored. Also, statistical analysis can define the most non-random words in the data set and use these

- 13 -

words as the vocabulary. Other statistical analysis could select the most appropriate words for the vocabulary.

In conjunction with or instead of the statistical analysis, the vocabulary can be reduced by utilizing certain n-gram words as equivalent. Proteins with similar function can diverge through evolution. In other words, evolutionary changes that involve conservative amino acids substitutions (replacing an amino acid with one of similar chemical character) often occur. Therefore, treating n-gram words representing amino acids having a similar chemical character as identical will further reduce the vocabulary without deleterious effects on the analysis. For example, the n-gram word GGE (glycine-glycine-glutamic acid) is quite different from the n-gram word GGD (glycine-glycine-aspartic acid). However, with a conservative amino acid substitution, the n-gram word for both will be equivalent, with the proteins having the different n-gram words being viewed as more similar. To reduce the size of very large vocabularies, even more latitude can be incorporated into the substitutions, for example, using a single letter for all polar amino acids. As the vocabulary is reduced using any of the above methods, longer n-gram words may be used. Thereby, comparisons of longer sequence fragments can be made.

After the vocabulary is created, a high-dimensional context vector is created for each sequence string, e.g. protein sequence, in the data set (step 340). To create the context vector, the n-gram words of each sequence string are compared to the vocabulary. In a binary scheme, the presence of an n-gram word will be used to place the magnitude of the context vector along the corresponding axis of the word in the

- 14 -

high-dimensional space at a predetermined amount. Alternatively, the number of occurrences of an n-gram word in the protein will be used to increase a magnitude of the context vector along the corresponding axis of the word in the high-dimensional space in proportion to the number of occurrences. The absence of a word will result in a zero value for the corresponding axis of the word in the high-dimensional space. Thereby, a context vector for each protein is the data set is created.

The use of n-gram words to create context vectors can also be used for sequences representing things other than protein. For example, nucleotide sequences have an alphabet of only four letters (G, A, T, C). When the alphabet is reduced, the word length of the n-gram can be increased.

2.   Predefined Domains

A second method for creating context vectors for biopolymer material uses predefined domains to define the attributes of the biopolymer material.

Generally speaking, proteins have evolved from a set of building blocks with each protein arising from a different combination of these building blocks. In proteins, building blocks are known as motifs, and represent structural or functional domains. All proteins are built from the same sets of motifs. Current research has identified approximately 5000 motifs so far.

In this method, predefined domains of interest in the protein data set, such as motifs, are selected (see Fig. 4, step 410). For example, a user could select a file including motif definitions from a public domain set, such as PROSITE (available at various locations including <http://www.expasy.ch/prosite>), or could select any

- 15 -

available file of motif definitions including a user-defined file. Because each motif is designated by all the combinations of sequences that could constitute that motif, each occurrence of the motif is designated with the same identifier. Thus, the method can account for degeneration of motifs. To reduce the number of selected motifs, combinations of motifs could be combined as a single motif. One method of combining motifs would be to treat the motifs as sequence letters and build n-gram words from contiguous structural attributes of the protein, as described above in connection with Fig. 3. Also, motifs can be combined based on their order of occurrence in the sequence.

Next, each of the selected predefined domains, e.g. motifs, is set to a respective axis in a high-dimensional space (step 420). After the axes are defined, a high-dimensional context vector is created for each protein in the data set (step 430). This is accomplished, for example, by determining the motifs of each protein. In a binary scheme, the presence of a motif will be used to place the magnitude of the context vector along the corresponding axis of the motif in the high-dimensional space at a predetermined amount. Alternatively, the number of occurrences of a motif in the protein will be used to increase a magnitude of the context vector along the corresponding axis of the motif in the high-dimensional space in proportion to the number of occurrences. Also, the methods for defining the magnitudes of axes could be combined, with motifs that indicate basic functionality represented as binary, and motifs whose functions depend on a number of occurrences represented in proportion to the number of occurrences. The absence of a motif will result in a zero value for the

- 16 -

corresponding axis of the motif in the high-dimensional space. Thereby, a context vector for each protein is the data set is created.

Since the resulting vectors can produce a sparse matrix, it may be necessary to couple the values representing motifs with other values, such as n-grams frequency. For example, the occurrence of n-grams can represent additional dimensions in the context vector (as described for step 420). Alternatively, an association matrix can be built showing the probability of co-occurrence of each n-gram with each motif. Most n-grams will have some probability of co-occurring with each motif, albeit small in some cases. The weight given each motif along its respective axis in the high-dimensional context vector can then be assigned by the sum of probabilities for each n-gram in that sequence having co-occurred with that motif.

3.   Geometric Shape

Referring to Fig. 5, another method for creating context vectors for biopolymer material will be described. In this method, the protein can be considered a geometric shape, where the complex characteristics and properties of that shape describe the protein function, instead of using motifs to define the entire protein structure. First, a collection of descriptors of characteristics and properties describes each protein in a manner analogous to that for motifs (step 510). Spline functions generate the surface by representing characteristics and features as variables. The descriptors are then collected for all proteins in the data set and each descriptor is defined as an axis for the high-dimensional context vector (step 520). The context vector is created for each protein (step 530). In a binary approach, the value along each axis is set to either one

- 17 -

if the descriptor is present or zero if absent. Like predefined domains, the value along each axis could alternatively be scaled by the number of occurrences for each descriptor.

4. Non-Geometric Indication

Fig. 6 is a flow chart of another method for creating context vectors corresponding to biopolymer material using an indirect, relative, and non-geometric indication of the structure or function of the biopolymer material, rather than an indicator of the actual structure or portion thereof (by, e.g., using n-gram words, motifs, or surfaces).

A common method used for analysis of biopolymer material, e.g., protein, involves comparing each protein's structure to a data set of proteins to determine each protein's similarity to each of the other proteins. One conventional method, the Basic Local Alignment Search Tool (BLAST), provides a list of proteins from the data set rank ordered by expect values. Various entities provide BLAST algorithms including <http://www3.ncbi.nlm.nih.gov/BLAST>. The BLAST method provides a probability score when comparing one sequence against another The score is usually expressed as the 'expectation' that in comparing the test sequence against a number of other sequences a match would have been found, e.g., a probability score of 1×10^{-180} would say that there was very little expectation to have found the match by chance and thus the two sequences must be related. A score of close to one indicates that the match would have been expected by chance, i.e., the homology is very weak. To provide the probability score, BLAST uses a heuristic algorithm that seeks local as opposed to

- 18 -

global alignments and is therefore able to detect relationships among sequences which share only isolated regions of similarity. In other words, the matches themselves are based on finding regions of similarity and, in the case of proteins, account for the fact that some amino acids could behave similarly at a given position. Thus, overall the BLAST method gives a very good picture of similarity and can be used to provide a non-geometric distance measure.

Specifically, BLAST segments a sequence string and searches for short regions that are identical. Each of these regions can serve as a nucleus for finding additional similarities, leading to an alignment between the sequences. The BLAST family of algorithms includes Gapped BLAST, which allows deletions and insertions in the alignments that are determined and provides scores potentially more reflective of biological relationships, and Position-Specific Iterated BLAST (PSI-BLAST), which derives initial BLAST scores and then uses those results to look for more distantly related sequences.

BLAST and the multitude of other methods for finding relationships (e.g., FASTA and Smith-Waterman) are able to find complex similarities including gaps between related regions.

When a large number of proteins are being analyzed, the list of output results can be voluminous and difficult to analyze, especially in a BLAST all-against-all experiment, where each item of biopolymer material, e.g. protein, from a user-identified set is compared to every other item in the set.

To create context vectors and, thus, more efficiently analyze a data set of a large number of items of biopolymer material, an all-against-all experiment is performed for

all of the biopolymer material in the data set using a method that determines a relationship between proteins (step 610). One such method determines a probability score based on whether mere chance can explain the similarity in composition of the protein.

The list output of the all-against-all experiment is read by a computer software routine that identifies the expect score for each comparison and arranges those as a square matrix in an electronic data file (step 620). An exemplary square matrix 700 representing a protein data set is shown in Fig. 7. Each protein indexes a respective row and a respective column of the matrix, so as to compare each protein with all other proteins. The values of the cells within the matrix are then populated to indicate how related the proteins are, that is, the expect score from the all-against-all experiment. In Fig. 7, illustrative circles 710 and 720 show that when protein P1 was compared to protein P2, an expect value of 3 was produced. Alternatively, similarity values could populate the matrix, where the similarity score would equal 1-expect score.

Once the matrix is populated, the rows or columns of the matrix are used to create a context vector for each protein in the data set (step 630). In one aspect of this implementation, this step is accomplished by considering an entire row or column as a vector. In other words, the vector for each of the N objects or proteins to be visualized, has N attributes. The i-th attribute is a comparison measurement between the i-th protein in the data set and the object protein. In Fig. 7, a context vector for the object protein PN would be $(1, 5, 8 \times 10^{-4}, \ldots 0)$, as shown by illustrative oval 730 or oval 740. All of the context vectors constitutes an object attribute matrix.

Where the method for deriving relationship of one protein to another provides additional information, that information can be appended to the matrix and then defined as categorical, numeric, textual, or sequence data where the data set is visualized. For example, PSI-BLAST can provide information on how many iterations were required to identify the protein as being related.

In another aspect of this implementation, the values of the matrix can be adjusted to eliminate distortions created by the similarity determination method. The adjustment can occur at any time prior to population of the matrix, as the matrix is populated, after the matrix is populated, or after the context vectors are created. For example, a BLAST score above a predetermined value is truncated to the predetermined value to eliminate the distortion of relationships that occurs with poor similarity, as a high BLAST score indicates low similarity. Also, a predetermined minimum acceptable BLAST score can be substituted for scores below the predetermined minimum acceptable BLAST score, so as to control the relationships defined by very high homology by accounting for rounding errors in the BLAST routine that result in zero expectation values. Similarly, the scale for the similarity can be adjusted, for example, by taking the log of the expect score, which can provide greater weight to the lower expect values.

Once context vectors are provided, the context vectors are projected onto a two- or three-dimensional viewing area (step 230 of Fig. 2). For example, the context vectors are used to create a distance matrix. Clustering may then determine a centroid for a subset of proteins, and then the clusters and objects, e.g. proteins, are projected

- 21 -

onto the two- or three-dimensional viewing area. Previously discussed U.S. Patent Application Serial No. _____, entitled DATA PROCESSING, ANALYSIS, AND VISUALIZATION SYSTEM FOR USE WITH DISPARATE DATA TYPES, describes one method of visualizing context vectors based on sequence data.

D.   Conclusion

While there has been illustrated and described what are at present considered to be a preferred implementation and method of the present invention, it will be understood by those skilled in the art that various changes and modifications may be made, and equivalents may be substituted for elements thereof without departing from the true scope of the invention.

Modifications may be made to adapt a particular element, technique, or implementation to the teachings of the present invention without departing from the spirit of the invention. For example, any living material, from organism to microbe, could be represented using the context vectors of the present invention. Further, the present invention is not limited to the biosciences, and any material or energy could also be represented.

Also, the foregoing description is based on a client-server architecture, but those skilled in the art will recognize that a peer-to-peer architecture may be used consistent with the invention. Moreover, although the described implementation includes software, the invention may be implemented as a combination of hardware and software or in hardware alone. Additionally, although aspects of the present invention are described

as being stored in memory, one skilled in the art will appreciate that these aspects can also be stored on other types of computer-readable media, such as secondary storage devices, like hard disks, floppy disks, or CD-ROM; a carrier wave from the Internet; or other forms of RAM or ROM.

Therefore, it is intended that this invention not be limited to the particular implementation and method disclosed herein, but that the invention include all implementations falling within the scope of the appended claims.

What is claimed is:

1.	A method for generating a high-dimensional vector representation of a sequence string of units in a data set including a plurality of sequence strings, comprising:

dividing each of respective sequence strings into blocks of at least three units to create a vocabulary of blocks;

defining a respective vector axis to correspond to blocks in the vocabulary;

determining for each vector axis whether a string includes a block corresponding to the respective vector axis; and

creating a high-dimensional vector based on the determination.

2.	The method according to claim 1, wherein the dividing step includes separating each of the respective strings into overlapping portions of three units or more.

3.	The method for generating a high-dimensional vector representation of an item of biopolymer material in a data set including a plurality of items of biopolymer materials, comprising:

identifying each item of biopolymer material by a respective series of at least three sequence letters, each sequence letter corresponding to a sub-item of the biopolymer material;

dividing each of the respective series of sequences letters into blocks of three sub-items or more to create a vocabulary of blocks;

defining a respective vector axis to correspond to blocks in the vocabulary;

- 24 -

determining for each vector axis whether a series of sequence letters representing an item of biopolymer material includes a block corresponding to the respective vector axis; and

creating a high-dimensional vector based on the determination.

4.   The method according to claim 3, wherein the step of dividing includes separating each of the respective sequences into overlapping portions of at least three sub-items.

5.   The method according to claim 3, wherein the biopolymer material is protein.

6.   The method according to claim 3, wherein the biopolymer material is nucleic acid.

7.   A method for generating a high-dimensional vector representation of an item of biopolymer material in a data set including a plurality of items of biopolymer material, comprising:

selecting predefined domains of the plurality of items of biopolymer materials;

defining a respective vector axis to correspond to the selected domains;

determining for each vector axis whether an item of biopolymer material includes a domain corresponding to the respective vector axis; and

creating a high-dimensional vector based on the determination.

8.   The method according to claim 7, wherein the predefined domains are motifs.

9.   The method according to claim 7, wherein the biopolymer material is protein.

10.   The method according to claim 7, wherein the biopolymer material is nucleic acid.

11.   A method for generating a high-dimensional vector representation of an item of biopolymer material in a data set including a plurality of items of biopolymer material, comprising:

defining each item of biopolymer material in the data set as a surface using descriptors of at least one of structure and function;

defining a respective vector axis to correspond to the descriptors;

determining for each vector axis whether an item of biopolymer material includes a descriptor corresponding to the respective vector axis; and

creating a high-dimensional vector based on the determination.

11.   The method according to claim 10, wherein the biopolymer material is protein.

12.   The method according to claim 10, wherein the biopolymer material is nucleic acid.

13.   A method for generating a high-dimensional vector representation of an item of

biopolymer material in a data set including a plurality of items of biopolymer material,

comprising:

comparing information regarding each biopolymer material of the plurality to

information regarding each other biopolymer material to provide a respective result;

arranging the results in a square matrix indexed by the plurality of items of

biopolymer materials;

creating a high-dimensional vector for an item of biopolymer material based on a

row or column of the matrix; and

creating a distance matrix based on the high-dimensional vector.

14.   The method according to claim 13, wherein from each row or column of the

matrix, a respective high-dimensional vector is created for each of the items of

biopolymer material based on the results in the row or column.

15.   The method according to claim 13, wherein the comparing uses a Basic Local

Alignment Search Tool.

16.   The method according to claim 13, wherein the comparing provides results

based on an expectancy.

17.   The method according to claim 13, wherein the biopolymer material is protein.

18.   The method according to claim 13, wherein the biopolymer material is nucleic

acid.

19.   An apparatus for generating a high-dimensional vector representation of an item

of biopolymer material in a data set including a plurality of items of biopolymer material,

comprising:

> at least one memory having program instructions, and

> at least one processor configured to execute the program instructions to

perform the operations of:

> > comparing information regarding each biopolymer material of the

plurality to information regarding each other biopolymer material to provide a respective

result;

> > arranging the results in a square matrix indexed by the plurality of

items of biopolymer materials;

> > creating a high-dimensional vector for an item of biopolymer

material based on a row or column of the matrix; and

> > creating a distance matrix based on the high-dimensional vector.

20.   An apparatus for generating a high-dimensional vector representation of an item

of biopolymer material in a data set including a plurality of items of biopolymer material,

comprising:

means for comparing information regarding each biopolymer material of the plurality to information regarding each other biopolymer material to provide a respective result;

means for arranging the results in a square matrix indexed by the plurality of items of biopolymer materials;

means for creating a high-dimensional vector for an item of biopolymer material based on a row or column of the matrix; and

means for creating a distance matrix based on the high-dimensional vector.

21.   A computer-readable medium containing instructions for controlling a computer system to perform a method for generating a high-dimensional vector representation of an item of biopolymer material in a data set including a plurality of items of biopolymer material, the method comprising:

comparing information regarding each biopolymer material of the plurality to information regarding each other biopolymer material to provide a respective result;

arranging the results in a square matrix indexed by the plurality of items of biopolymer materials;

creating a high-dimensional vector for an item of biopolymer material based on a row or column of the matrix; and

creating a distance matrix based on the high-dimensional vector.

ABSTRACT OF THE DISCLOSURE

Systems for creating high-dimensional vectors representing sequence strings and biopolymer materials are provided. A first system for divides respective sequence strings into blocks of at least three units to create a vocabulary of blocks. A second system selects predefined domains of a plurality of items of biopolymer materials. A third system defines each item of biopolymer material in a data set of biopolymer materials as a surface using descriptors of at least one of structure and function. A fourth system compares information regarding each biopolymer material of a plurality of biopolymer materials to information regarding each other biopolymer material.

Memory

Computer
Process

110

104

Processor

102

100

Router

125

To Network

106

132

134

124

122

130

120

FIG. 1

Access Data Set ⌇～ 210

Create Context Vector ⌇～ 220

Visualize Data Set ⌇～ 230

FIG. 2

Identify Sequence ~ 310

Create Vocabulary ~ 320

Define Axes ~ 330

Create Context Vector ~ 340

FIG. 3

```
┌─────────────────────┐
│   Select Predefined │  ⌇ 410
│      Domains        │
└─────────────────────┘
          │
          ▼
┌─────────────────────┐
│    Define Axes      │  ⌇ 420
└─────────────────────┘
          │
          ▼
┌─────────────────────┐
│ Create Context Vector│ ⌇ 430
└─────────────────────┘
```

FIG. 4

```
┌─────────────────────────┐
│                         │
│     Define Surface      │⌇~ 510
│                         │
└─────────────────────────┘
             │
             ▼
┌─────────────────────────┐
│                         │
│      Define Axes        │⌇~ 520
│                         │
└─────────────────────────┘
             │
             ▼
┌─────────────────────────┐
│                         │
│   Create Context Vector │⌇~ 530
│                         │
└─────────────────────────┘
```

FIG. 5

Compare All vs. All — 610

Arrange Matrix — 620

Create Vector — 630

FIG. 6

FIG. 7